Synthese (2006) 153:377–391
DOI 10.1007/s11229-006-9098-9

ORIGINAL PAPER

Doing cognitive neuroscience: a third way

Frances Egan · Robert J. Matthews

Received: 6 July 2006 / Accepted: 8 August 2006 /
Published online: 20 October 2006

Abstract The "top-down" and "bottom-up" approaches have been thought to exhaust the possibilities for doing cognitive neuroscience. We argue that neither approach is likely to succeed in providing a theory that enables us to understand how cognition is achieved in biological creatures like ourselves. We consider a promising third way of doing cognitive neuroscience, what might be called the "neural dynamic systems" approach, that construes cognitive neuroscience as an autonomous explanatory endeavor, aiming to characterize in its own terms the states and processes responsible for brain-based cognition. We sketch the basic motivation for the approach, describe a particular version of the approach, so-called 'Dynamic Causal Modeling' (DCM), and consider a concrete example of DCM. This third way, we argue, has the potential to avoid the problems that afflict the other two approaches.

Keywords Neuroscientific cognitive modelling · Top-down approach to neuroscience · Bottom-up approach to neuroscience · Neural dynamic systems · Dynamic causal modeling · Neural structural–functional relations · Neural connectivity

1 Introduction

The goal of cognitive neuroscience is to explain how cognitive processes emerge from neural activity. To hear many philosophers of mind and cognitive psychologists tell it, there are basically only two ways of doing cognitive neuroscience, either top-down or bottom-up, and of these two ways only the top-down approach has even the ghost

F. Egan (✉) · R. J. Matthews
Department of Philosophy, Rutgers University,
126 Nichol Ave.,
New Brunswick, NJ 08901, USA
e-mail: fegan@rci.rutgers.edu

of a chance of success. As the bottom-up approach is usually conceived, or at least as it is caricatured, it is indeed hopeless. But the confidence of many philosophers and cognitive psychologists in the top-down approach is also misplaced, or so we will argue. So if these two approaches exhaust the field of possibilities, then cognitive neuroscience, and cognitive science more generally, is in deep trouble, because it seems unlikely that either will succeed in providing what all cognitive scientists want, namely, a theory of cognition that is ultimately anchored in brain processes, a theory that will enable us to understand how cognition is achieved in biological creatures like us.

There is in fact a promising third way of doing cognitive neuroscience, one that seemingly avoids the problems that afflict top-down and bottom-up approaches. But before describing this third way and its potential virtues, let us first describe briefly the two ways of doing cognitive neuroscience that are commonly thought to exhaust the possibilities.

What is really at issue between these two approaches is the relative autonomy of neuroscientific explanations of cognition and cognitive phenomena: the bottom-up strategy presumes the possibility of a completely *sui generis* neuroscientific explanation of cognition, whereas the top-down approach presumes that neuroscience can at best provide an implementation story for cognitive psychological explanations, inasmuch as the real explanatory work, its proponents claim, is done at the cognitive, rather than the neural, level.

2 The bottom-up approach

The bottom-up approach, as the name suggests, presumes that neuroscientists investigating small scale neural phenomena will eventually be able to extrapolate from their accounts of small scale processes to the level of molar cognitive processes, with these molar processes emerging out of the small scale processes. The bottom-up slogan is something like "single cell processes today; tomorrow the entire brain". This optimistic view is expressed in Barlow's (1972) claim:

> A description of [the] activity of a single nerve cell which is transmitted to and influences other nerve cells and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. There is nothing else "looking at" or controlling this activity, which therefore must provide a basis for understanding how the brain controls behavior. (p. 380)

But as cognitive theorists of many stripes have noted, the early enthusiasm of brain researchers in single cell recordings has not resulted in an understanding of more complex brain processes, much less of cognitive processes:

> Modern neurophysiology has learned much about the operation of the individual nerve cell, but unpleasantly little about the meaning of the circuits they compose in the brain. The reason for this can be attributed, at least in part, to a failure to recognize what it means to understand a complex information-processing system; for a complex system cannot be understood as a simple extrapolation from the properties of its elementary constituents. (Marr & Nishihara, 1978, p. 28)

As Marr later put it,

> . . . trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers. It simply cannot be done. (Marr, 1982, p. 27)

The persistent complaints of computational theorists such as Marr seem borne out, that merely mucking around with low-level detail is unlikely to eventuate in any understanding of complex cognitive processes, for the simple reason that bottom-up theorists don't know what they are looking for and thus are quite unlikely to find anything. Sometimes the problem here is presented as a search problem: How likely it is that one can find what one is looking for, given the complexity of the brain, unless one already knows, or at least has some idea, what one is looking for. At other times the problem is presented less as an epistemological problem than as a metaphysical problem: cognitive processes, it is claimed, are genuinely emergent out of the low-level processes, such that no amount of understanding of the low-level processes will lead to an understanding of the complex cognitive processes.

3 The top-down approach

The top-down approach, as the name implies, presumes that one should begin with a well-developed psychological theory, perhaps computational in nature, and then look for the manner in which the various states and processes postulated by the theory are implemented. On this account, the explanatory work of the theory is done chiefly by the psychology and all that the neuroscience contributes is the implementation story, explaining how, in biological creatures like us, the hypothesized states and processes are realized.

Commenting on his own 1974 work on color vision, Marr says approvingly,

> . . . gone is any explanation *in terms of* neurons—except as a way of implementing a method. And present is a clear understanding of what is to be computed, how it is to be done, the physical assumptions on which the method is based, and some kind of analysis of algorithms that are capable of carrying it out. (Marr, 1982, p. 18, emphasis in original)

The approach is described as "top-down" inasmuch as it presumes a hierarchy of different levels of explanation, with each level in the hierarchy levying constraints on the next level below. For example, in Marr's own (1982) account of the explanatory hierarchy, the theory of computation—the specification of the cognitive function computed by the system—levies constraints on the level of representations and algorithms, and the level of representations and algorithms on the neural implementation story.[1]

The top-down approach to cognitive theorizing is nicely illustrated by Gallistel's (1990) discussion of the dance of the foraging bee. According to Frisch (1967), foraging bees returning from a nectar source perform a dance that specifies the direction and distance of the source. A specific segment of the dance—the "waggle"—represents, by its angle with respect to the vertical, the direction of the nectar source

[1] In practice, though, the top-down methodology often breaks down. Marr (1982), for example, appeals to neurological considerations to constrain the choice of algorithms.

relative to the sun. The number of waggles represents the distance to the source. Gallistel comments:

> There are no implementational mysteries here, if we grant the bee's nervous system the functional architecture [of a Turing machine]. When the bee is at the source, its dead-reckoning vector specifies the source's location relative to the hive (a vector structure). As the bee ingests the nectar preparatory to carrying it back to the hive, it gets a sensory input specifying its richness (a scalar structure). It writes the location vector and the richness scalar to memory. When it reaches the hive, the dance program retrieves these data structures (reads the memory). (2006, p. 67)

Here we see the top-down strategy in action. The bee's ability to convey information about the direction, distance, and quality of a nectar source by producing a dance with certain structural properties, Gallistel claims, constrains its neural architecture. It needs both a read/write memory that persists sufficiently through time and can be accessed by the bee's dance program and data structures that represent the distance, direction, and quality of the nectar. In other words, the bee needs a classical "rules and representations" architecture. A (connectionist) network architecture, Gallistel claims, is simply unable to account for the bee's behavioral capacity:

> There is no way to account for the behavioral facts except by assuming a read/write memory and the composition of data structures, both of which functionalities are absent in [a network] architecture. Therefore, we must assume the existence of mechanisms in the nervous system that perform these functions, despite the fact that what we currently know about the nervous system does not support such an assumption. On this analysis, what we know about behavior is a guide to what we must look for in the nervous system. (p. 65)

The implications for cognitive neuroscience could not be clearer: the brain has to be an information-processor with a "rules and representations" computational architecture.

Gallistel finds analogies from the history of other sciences to support his claim that high-level descriptions of phenomena determine the course of theorizing about implementing mechanisms. Lord Kelvin, unaware of radioactivity, pointed out to Darwin that the enormous age of the earth implied by the theory of evolution was not consistent with physicists' accounts of the possible duration of heat-generation processes in a body the size of the sun (Burchfield, 1990). Prior to Watson and Crick's discovery of the molecular structure of DNA, many biochemists found the concept of a gene unintelligible (Judson, 1980). They simply had no inkling of a molecular structure that could both make a copy of itself and determine the sequence of amino acids in the synthesis of a protein. Kelvin and the biochemists, as Gallistel puts it, "did not know what they did not know." (p. 70) The moral Gallistel draws from these historical examples is that neuroscientists should look to cognitive psychologists to tell them about the structure of the brain:

> If behavior is the last court of appeal, then there are mechanisms in the nervous system not yet dreamed of in the philosophy of neuroscientists. (p. 70)

Gallistel's confidence that behavioral evidence alone can tell us about neural mechanisms is misplaced. The finite domains over which neural computations are defined insure that whatever the functions these mechanisms compute, they can be computed by connectionist architectures. Indeed, it is provable that standard multilayer feedforward connectionist architectures can approximate, to any desired degree of accuracy,

any function computable by means of a classical architecture (Hornik et al, 1989), so Gallistel is presumably not claiming that a connectionist architecture cannot accomplish the computational feats that are constitutive of the bee's cognitive capacity. It is, to be sure, an open question whether any given computational architecture can accomplish these feats with the right sort of complexity profile. Considerations of computational complexity might in principle decide this question, and thus decide among alternative neural computational architectures. But it is notoriously difficult to bring behavioral evidence to bear on complexity considerations, and here, too, behavioral evidence alone is not decisive.

What Gallistel is presumably claiming must then be something about how the bee's cognitive capacity is to be explained. Presumably he is claiming that this capacity must be explained in terms of certain distally interpreted representations and certain computations defined over these representations, not because only certain computational architectures are capable of computing whatever functions are constitutive of the capacity in question, but because for certain unspecified epistemological reasons only certain kinds of computational explanations will count as genuinely explanatory of this capacity. But if this is what Gallistel is claiming, then one must wonder how he can be so confident that he is not laboring under just the lack of knowledge that he attributes to Kelvin and the biochemists. In Gallistel's case, the lack of knowledge could have to do, with how the bee's brain manages to do, without a "rules and representations" architecture, what we theorists do using such an architecture. Just as the resolution of the above inter-theoretic tensions turned on revolutionary discoveries of radioactivity and the molecular structure of DNA—discoveries at the level of implementing mechanisms—so a full understanding of how the brain implements behavioral and cognitive capacities may depend on future developments in functional neuroscience. And just as surely, nothing requires that scientific explanations take the form that we think that they should take. The proper form of scientific explanation for a domain cannot be given apriori; it must flow, in part, from the discovered nature of the phenomenon being explained.

What *is* true in what Gallistel says is that an account of the bee's capacity in terms of a classical "rules and representations" architecture renders the capacity *explanatorily transparent* in a way that an account in terms of a connectionist network architecture would not, by depicting in a perspicuous fashion the flow of information through the system. The classical architecture posits data structures that *explicitly represent* precisely the information that the cognitive psychological account attributes to the bees, viz., information about the direction, distance, and quality of the nectar source. And it provides a computational account of how this represented information is used by the bee in its waggle dance. So if a cognitive theorist were to set out to build a device that replicates the bee's capacity, with no concerns about available hardware or physiological constraints, she might naturally be inclined to do so in a classical architecture. But, of course, the fact that a classical architecture is the methodologically tractable solution for the theorist does not imply that it is the brain's solution. There is no reason to suppose that Nature values explanatory transparency.

The general difficulty with the top-down approach becomes apparent when one considers the sort of relations that must hold between levels if one level is to levy constraints on the next lower level: one is going to need some sort of mapping relation that takes the states, processes, and entities of one level into those of the other. A nice picture to be sure, but it is its great weakness as well, for to the extent that one

cannot effect the mapping, one lacks any way of imposing the constraints, and hence of developing the envisioned implementation story.

The recipe that self-styled "cognitivists" propose for effecting the mapping requires the theorist to characterize at the top-most level a rich cognitive structure of intentionally interpreted representational internal states ('intentional internals', for short).[2] The cognitive capacity to be explained—e.g., recovering the three-dimensional structure of the scene, recognizing faces, understanding speech—is typically decomposed into a series of subtasks, each of which is itself characterized in intentional terms. The intentional internals posited by the cognitive theory are presumed to be distally interpretable, i.e., to represent such external objects and properties as the orientation of surfaces, facial features, spatial locations, etc. It is thought that if these intentional internals are not distally interpretable, then the account is unlikely to yield an explanation of the organism's successful interactions with its environment. Moreover, cognitive processes must preserve certain epistemic and semantic relations defined over these representations. The outputs of these processes should *make sense*, should be *rational*, given the inputs. This rich cognitive structure constrains theorizing at the lower levels. Cognitive theorists then look for computational and neural states to realize the intentional internals. The outcome, if things go well, will be a mapping between the causal structure of the mind and the causal structure of the brain.

But why think that things will typically go well? We can identify in the cognitivist picture two constraints that regulate theorizing at the topmost level:[3] (1) the *intelligibility constraint*, which requires that later internal states (and behavioral outputs, if any) make sense, are rational, given earlier states (and inputs, if any); and (2) the *distal interpretability constraint*, which requires that the posited internal states represent external objects and properties. Theory construction in neuroscience is not, at least not in the first instance, governed by these constraints. Neuroscientists, for example, first identify various spike trains, and only then, if they are pursuing a top-down approach, undertake to discover what, if any, environmental variables these spike trains correlate with. The point here is that neuroscience, like every science, has its own explanatory constraints governing the individuation of the states, events, and processes within its domain. And even if one ultimately tries to construct a mapping of psychological states, events, and processes into these neural states, events, and processes, one begins with the ontological and taxonomic commitments indigenous to neuroscience.[4] The worry about the top-down approach, then, is this: why think that there will be intermediate steps at the level of neural processing corresponding to the processes posited at the cognitive level, or neural states corresponding to the intentional internals? There would seem to be no apriori reason to think that the brain works this way, and hence neuroscience, which is governed by quite different

[2] See, for example, Haugeland, 1978 ("... the fundamental idea of cognitive psychology [is that] intelligent behavior is to be explained by appeal to internal "cognitive processes"—meaning, essentially, processes interpretable as working out a rationale." pp. 260–261), and Marr, 1982 ("Modern representational theories conceive of the mind as having access to systems of internal representations; mental states are characterized by asserting what the internal representations currently specify, and mental processes by how such internal representations are obtained and how they interact." p. 6).

[3] There are no doubt other constraints.

[4] Even strict top-down theorists are not committed to the view that the same set of explanatory constraints governs theorizing at all levels in the hierarchy. The idea that the structure of entities and processes posited at higher levels constrains the search for implementing mechanisms at lower levels does not require this strong "unity of science" thesis.

explanatory constraints, need not assume that it does. But the top-down project's success depends precisely on finding neurologically individuated states and entities that correspond systematically to the taxonomy imposed from the top. Of course, the fondest hopes of the top-down cognitivist might be borne out. The relations between the levels in the hierarchy might, *per mirabile*, turn out to be explanatorily transparent in just this way. But despite the optimism of Gallistel and other top-down cognitivists, we see no reason to think that they will. It is just as likely that the two taxonomies will cross-classify, and that the relation between the levels will be explanatorily opaque.[5]

Now, when pressed about the feasibility of their approach, top-downers typically reply with some version of the "how else could it be done?" gambit. They concede their reductionist commitments, but say in effect that if these commitments turn out to be false, then cognitive neuroscience, and indeed the entire project of explaining how biological creatures like us are capable of cognition, is hopeless. Top-downers might emphasize that they are not reductionists by choice: it's a gambit that the possibility of success requires.

4 A third way: a neural dynamic systems approach

The top-down strategy basically involves trying to map cognitive function, as defined by a (computational) cognitive theory of some cognitive capacity, into an independently specified neural structure. And in so doing, it assumes that the taxonomy of the cognitive theory is sufficiently like the taxonomy of neural structure that the one can be mapped into the other, maybe not perfectly, but good enough to enable one to get a handle on the neurology, such that one can then tweak the cognitive psychology sufficiently to bring the latter into synch with the former.

In their more sober moments, top-downers worry about whether cognitive and neural models are sufficiently well-specified to permit the construction of a mapping. A further worry, of course, is whether currently available cognitive and neural models are *correct*. Some cognitive theories (in psycholinguistics and early vision, for example) are specified in precise mathematical terms, yielding fairly precise predictions of behavior; but others consist primarily of vague descriptions not far removed from their roots in commonsense psychology. It is quite likely that much of current cognitive psychology is just plain wrong, at least in the details—what else would one expect of a relatively young discipline?—and so thoughts of a systematic mapping of fully elaborated cognitive function to neuroscience may be very premature.

In light of these concerns, Poldrack et al. (2006) propose that cognitive neuroscience take as its goal not the systematic mapping of cognitive function into neural structure, but the specification of mapping relations between the "ontologies" that, as they put it, "specify the structures of cognitive function and neural systems, respectively" (p. 2).[6] They explain their proposal as follows:

[5] Readers may recognize a similarity between the argument advanced here and Davidson's argument for the anomalousness of the mental, but with an important difference. Davidson appeals to the constitutive ideal of rationality governing the ascription of intentional states to argue against the possibility of mind-brain reduction. The argument advanced here is overtly *epistemological*. The cognitivist's account of mental processes might smoothly reduce to neuroscience. But there is no reason to believe that it will.

[6] This use of "ontology" is from the field of bioinformatics. See Bard and Rhee (2004) for a review of the relevant literature.

> Although in the end it would be desirable to map between fully specified causal models, given the current state of knowledge it may be useful to describe these structures using the more limited tools of ontologies. The expression of theories in terms of ontologies would help ensure that cognitive theories are specified to a degree that the relations between cognitive and neural processes can be formally characterized. (pp. 2–3)

The mapping of cognitive to neural ontologies is suggested, presumably, because it is thought to be a more modest and hence more easily achievable goal for cognitive neuroscientists to pursue. Ontologies do not presuppose the rich theoretical structure that would constrain a mapping between fully specified models. Or so it seems. In fact, it is hard to know what to make of them. As described by Poldrack et al. ontologies are something of a grab-bag, comprising "formalized knowledge bases that describe the structure of a specific domain" (2). Their structures vary from "controlled vocabularies (which outline a set of terms) to simple taxonomies which describe parent–child relations, to fully specified logical systems describing complex relations between entities" (2). Indeed "ontology" seems to include just about any theoretical commitment that any psychological theory might come up with.[7] But if ontologies do not presuppose the rich theoretical structure that would constrain a mapping between fully specified models, then one has to wonder about the empirical significance of any mapping in which they figure. After all, mappings are cheap, i.e., easily come by, if they are not sufficiently constrained. Would such a mapping between ontologies license any predictions and explanations about the entities in the domain of either framework? More generally, in the context of concerns about the specificity and correctness of cognitive and neural models, the move from fully specified theoretical structures to ontologies is not promising. The proposal looks to be little more than a desperate casting about for a mapping from the vague and confused to the probably wrong!

So what is the alternative? There is a third approach, a middle way that is neither bottom-up nor top-down, but that construes cognitive neuroscience as a largely autonomous explanatory endeavor, aiming to characterize in its own terms the states and processes responsible for brain-based cognition. This third way is what might be called the *neural dynamic systems* approach.[8] We shall sketch the basic motivation for the approach, describe a particular version of the approach, so-called "Dynamic Causal Modelling" (DCM), and then consider a concrete example of DCM.

Kiebel, Stephan, and Friston (2006) note that neuroscientists have had trouble specifying Structural–Functional Relationships (SFRs) for anatomically defined cortical areas of the brain, despite the fact that the anatomical microstructure of much of the cortex is well-understood. SFR hypotheses have the following form: "The brain component C has the functional property F because of its structural property S" (p. 3), where the brain components in question are gross anatomical areas such as Brodmann areas (V1, V5, etc.), Broca's area, and the inferior fusiform gyrus (IFG), and the functional properties are neurologically or cognitively defined functions such

[7] See, for example, the ontology included as an appendix to the DARPA solicitation, referenced in Poldrack et al. (2006): http://www.darpa.mil/ipto/solicitations/open/05–18_PIP.htm

[8] There is a growing literature on dynamic systems as alternative to cognitive *computational* models (see, e.g., van Gelder, 1997). The neural dynamic systems approach considered here does not cast itself as such an alternative.

as color vision, attention, etc. SFR hypotheses posit for some neural structure an explicitly *causal* role in some neurological or cognitive function.

There are several reasons why SFRs for cortical areas have not been forthcoming: (1) cortical areas are typically involved in more than one function;[9] (2) interactions between multiple structural variables seem to be involved in their operation; and (3) functional responses in the cortex are highly context-sensitive, depending on processing history as well as inputs from other brain areas. Examples of (3) are particularly noteworthy. Responses in visual areas can be dramatically altered by changes in attention, or other aspects of so-called "cognitive set" (Li, Piech, & Gilbert, 2004; Luck, Chelazzi, Hillyard, & Desimone, 1997). Even more striking are "paradoxical lesion effects", where a cognitive function is disrupted after a first lesion, and then restored after a second lesion (Lomber, Payne, Hildetag, & Rushmore, 2002; Sprague, 1966)! Kiebel et al. argue that these phenomena suggest the need for a dynamical account of neural processes, given that the role that a particular neural structure plays in cognitive function turns out to depend on the dynamical, time-sensitive interaction of this structure with other structures. Finally, (4) no cortical area operates in isolation but is connected to many other areas by anatomical long-range connections ("association fibers"). The upshot is that the behavior of a particular area cannot be predicted and explained from local microstructure alone. It has become increasingly clear, Kiebel et al. conclude, that cortical function will not be understood without adopting an explicit *dynamical systems* perspective.

It is worth dwelling on the implications of the brain's "connectivity" for the top-down strategy, which depends on the realistic possibility of identifying neural structures that realize the states and entities characterized functionally at the cognitive level. It seems clear from the evidence that the function subserved by a particular cortical structure at a given time depends on aspects of the "cognitive set" (including attention), on activity in the areas to which the structure is linked by anatomical long-range connections, and probably on a whole host of other "connectivity effects". Structure–Function Relations (SFRs)—or even the weaker Structure-Function Correlations (SFCs), which imply no causal connection—cannot be specified except in a context that takes into account the connectivity of the structure in question with other cortical structures. Moreover, it is clear that the causal potential of neural structures changes over time. Recall paradoxical lesion effects—the restoration of lost cognitive function after a second lesion. So the prospect of finding neural structures that implement specific functional states characterized independently at a higher level of theory looks bleak. In addition, given the apparently dynamical character of neural processes, the fundamentally non-dynamical character of the structures characterized at the cognitive level (both the representational data structures over which cognitive processes are defined as well as the processes themselves are decidedly static in character) makes the prospects of finding a mapping of cognitive function into neural structures all the more bleak.[10]

Recent advances in brain imaging have made apparent the need for more sophisticated analytical tools for studying the brain. The standard methods available, until recently, for establishing the involvement of a cortical area in some cognitive

[9] For example, Broca's area has been shown to be involved in language processing, action observation, and local visual search (Hamzei et al., 2003; Manjaly et al., 2003).

[10] The initial stages of the investigation are likely to yield spurious mappings, given that cortical areas are typically involved in multiple cognitive functions. Whether a structure plays a role in a given cognitive function at a time depends on aspects of the "cognitive set".

function—invasive recordings from animals and lesion studies on human subjects—failed to provide much more than the grossest Structure–Function Correlations (SFCs). Positron emission tomography (PET), developed in the 1980 s, and functional magnetic resonance imaging (fMRI), developed in the early 1990 s, made possible high resolution measurements of area-specific changes of brain activity that are correlated with components of cognitive tasks.[11] Since the introduction of these imaging techniques there has been an explosion in the number of proposed SF correlations. Still, Kiebel et al. argue, there has been little progress toward an understanding of the causal mechanisms underlying the observed correlations. They conclude:

> To provide us with a deeper understanding of SFRs in the brain, functional neuroimaging will need to adopt an *explicit systems perspective*, using causal models of brain functioning that are based on neuroanatomical information about the structure of the investigated system, particularly with regard to the connectivity between areas. (p. 5, emphasis added)

The neural dynamic systems approach is based on the idea that complex systems can only be understood by finding a mathematical characterization of how their behavior emerges from the interaction of their parts over time. The success of dynamical approaches in other sciences, e.g., physics, biology, ecology, economics, etc., suggests that a similar approach in neuroscience might pay. Adopting a dynamical systems perspective in neuroscience, Kiebel et al. argue, makes possible powerful new insights about SFRs (Structure–Functional Relationships), and hence an understanding of how function depends (causally) on brain structure.[12] In the first place, dynamic systems analyses provide the basis for precise formal definitions of *structure*, *function*, and *SFR*.[13] Second, they allow for the expression of SFRs in quantitative terms, which makes possible quantitatively precise predictions of system behavior. Third, and most importantly, Kiebel et al. claim that a formal dynamical systems analysis "is the only way to fully *understand* how a system works" (p. 7, emphasis in original).

5 An illustration: dynamic causal modeling (DCM)

So what would a neural dynamic systems approach to cognitive neuroscience look like? There are a number of different possibilities, but recent work in so-called "dynamic causal modeling" (DCM) offers a useful illustration.[14] DCM undertakes to construct a realistic model of neural function, by modeling the causal interaction among anatomically defined cortical regions (such as various Brodmann's areas, the inferior temporal fusiform gyrus, Broca's area, etc.), based largely on fMRI data. The idea is to develop a time-dependent dynamic model of the activation of the cortical regions implicated in specific cognitive tasks. In effect, DCM views the brain as

[11] PET measures changes in cerebral blood flow (rCBF); fMRI measures blood oxygen-level dependent (BOLD) signals.

[12] Recall that SFRs, unlike SRCs (Structure–Function Correlations) specify causal dependencies.

[13] A *system* can be defined informally as "a set of elements which interact with each other in a spatially and temporally specific fashion." (p. 7) *Structure* refers collectively to the static, time-invariant, properties and relations of the system, and *function* to the dynamic, time-variant properties and relations which depend on structure. The *structure–function relationship* (SFR), then, is defined by the nature of this dependence. These informal definitions are given mathematical form using a set of differential equations with time-invariant parameters.

[14] For a general discussion of DCM, see Friston (2003) and Friston et al. (2003).

a dynamic system, consisting of a set of structures that interact causally with each other in a spatially and temporally specific fashion. These structures undergo certain time-dependent dynamic changes in activation and connectivity in response to perturbation by sensory stimuli, where the precise character of these changes is sensitive not only to these sensory perturbations but also to specific non-stimulus contextual inputs such as attention. DCM describes the dynamics of these changes by means of a set of dynamical equations, analogous to the dynamical equations that describe the dynamic behavior of other physical systems such as fluids in response to perturbations of these systems.

The fMRI methodology employed by DCM to measure systems dynamics is not of particular importance here. What we want to focus on is rather the sort of model of neural function that DCM offers and its relation to the more traditional bottom-up and top-down approaches.

Let us consider, as an example of DCM modeling, studies by Buchel and Friston (1997) and Friston and Buchel (2000) of the modulating effects of attention on connectivity between the superior parietal cortex (SPC) and visual area V5 and between the inferior fusiform gyrus (IFG) and SPC. In these companion studies, subjects viewed optic flow stimuli comprising radially moving dots, moving out from a fixed fixation point at a constant velocity. In some trials subjects were asked to detect possible changes in velocity (which did not actually occur). The sensory perturbatory input was modeled by photic stimulation of V1, while the specified motion of the dots was modeled by a non-stimulus contextual modulation of V1–V5 connectivity. Attention was modeled by a non-stimulus contextual modulation of connectivity between IFG and SPC and between SPC and V5. fMRI studies were then undertaken to determine the affective connectivity between these regions based upon the hypothesized inputs, largely confirming predicted modulating effects of attention.

What is attractive about neural dynamic systems approaches such as DCM is the possibility that through causal modeling of various cognitive tasks neuroscience might develop a set of simultaneous dynamical equations that, like the dynamic equations for other physical systems (e.g., the Navier–Stokes equations for fluids), would enable us to predict the dynamic behavior of the brain under different sensory perturbations and various non-stimulatory contextual inputs (such as attention). Obviously, neuroscience is still a *very*, *very* long way away from such an eventuality. But arguably no further away than are the two traditional approaches to cognitive neuroscience to characterizing brain-based cognition. And we do in fact know something about how to go about developing such a neural dynamic account of cognition, given our success in developing dynamical theories for other sorts of physical systems.

If cognitive neuroscience were to be successful in developing such dynamical equations that describe the behavior of the brain in the course of various cognitive tasks, then what many take to be the Holy Grail of cognitive neuroscience, namely specifying structure–function relationships, might be within reach. For once the dynamical equations were in hand, then these equations could in principle[15] be solved (by computational methods, to be sure) to explain how neural dynamics results from neural structure.[16] Of course, the structure-function relationships in question will not

[15] We say "in principle" because the solution of these equations, like the solutions for the dynamical equations for other physical systems, is likely to be computationally intractable if neural systems implicated in various cognitive tasks are not reasonably simple, as measured by the number of structural elements and inputs.

[16] For discussion, see Kiebel et al. (2006).

necessarily be those that so-called "cognitivists" have in mind, since neither structure nor function will be of the sort they have in mind. But it will be a kind of structure–function relationship that will do what we expect an explanation of cognition to do, namely explain how cognition is possible in biological creatures like us.

6 General discussion

Let us summarize, briefly, some implications of the neural dynamics systems approach, as illustrated by DCM, for the problem of explaining cognition. Notice first, and most importantly, it explains cognition in the sense that it gives an account of the phenomena in neural terms, emphasizing a break with the "cognitivist" assumption that to be an *explanation of cognition* is necessarily to be a *cognitive explanation*, where by the latter one means an explanation that traffics in the usual semantic and intentional internals dear to cognitivists. Second, the account leaves open the question of whether there is any interesting mapping between cognitive explanations and neural explanations of cognitive phenomena—maybe there is, but then again maybe not. Third, the approach makes no effort to satisfy the constraints governing cognitivist theorizing.

We will conclude by considering some objections to the proposed "third way".

6.1 Objection #1

Isn't this "third way" really just a *bottom-up* approach?

6.1.1 Reply

As noted earlier, the general rap by top-downers against the small scale investigation characteristic of the bottom-up approach is that theorists have their heads hopelessly lost in the neural details, with no idea of what they are looking for, and hence are unlikely to end up with anything remotely *cognitive*. This is a caricature of much neuroscientific research. In any event, DCM theorists are not trying to extrapolate the behavior of cortical structures from individual neurons, or even from assemblies of neurons. The neural dynamic systems approach, at least as illustrated by DCM, seeks to uncover the causal role of large scale neural structures in neural function, so the approach really does occupy a middle ground between the other approaches.

There is actually more commonality between DCM theorists and cognitivist top-downers than one might imagine. Both begin with functional accounts of the processes that underlie the cognitive phenomena in question. And both then undertake to ascertain the structures that realize these functions. But there is this important difference: the cognitivist approach spells out the functional account in intentional terms, attributing processes defined over contentful states, whereas DCM does not. Moreover, DCM does *not* aim to discover neural structures that implement intentionally characterized states.

6.2 Objection #2

Right, and that's exactly what is wrong with DCM! Dynamic Causal Models, and neural dynamic approaches more generally, don't provide *cognitive* explanations.

Genuinely cognitive explanations respect what Bickle has called "the standard mark of the cognitive" (1998, p. 166); they recognize the explanatory need for representations and computations over their contents.[17]

6.2.1 Reply

Yes, DCM does not provide cognitive explanations, by this criterion. It doesn't posit "intentional internals" and processes defined over them. As noted above, DCM investigation is not regulated by the constraints, most notably *intelligibility* and *distal interpretability*, that govern "cognitivist" theorizing. But what follows from this? All that follows, we suggest, is a need to distinguish *cognitivist explanations* from *explanations of cognition*. DCM seeks to explain the phenomena that commonsense, and perhaps cognitive psychology, take to be cognitive. But in taking these phenomena as *explananda*, DCM recognizes no implication regarding the proper form of explanation of these phenomena. The situation is perhaps analogous to disorders that are specified as psychological because of the sort of explanations that are initially offered—typically, shallow "folk" explanations—even though the neuroscientific explanation of the disorder may not traffic in any psychological notions.[18] It is an interesting question just what defines the domain of cognitive phenomena, but we do not have to answer that question here.

6.3 Objection #3

The real problem is that DCM models, and neural dynamic systems approaches more generally, are not really *explanatory* of cognition. Like all purely neural accounts of cognition, they fail to explain *why* the outputs of cognitive processes are rational given the inputs. Only generalizations adverting to mental representations and processes defined over them can do that.

6.3.1 Reply

Understanding cognition requires that we understand how sensory information plus current brain states determine behavior and other brain states. If behavioral outputs are in fact rational, given the inputs, then understanding cognition would require explaining why that should be so, and DCM, and neural dynamic systems approach more generally, would be committed to providing such an explanation. But so far we have been given no reason to suppose that such an explanation would have to advert to intentional internals. There is, after all, no independent evidence that the brain actually *has* intentional internals, beyond, of course, our commonsense way of characterizing the mental causes of behavior, but commonsense is hardly dispositive.

Moreover, and perhaps more to the point of the present objection, the appearance that positing intentional internals is by itself explanatory of cognition is illusory. There is certainly a heuristic payoff for positing intentional internals—they allow us to keep

[17] As Bickle (1998) puts it: "We can only explain genuinely cognitive psychological phenomena by generalizations adverting to the contents of mental representations and to computations over these contents." (p. 2)

[18] Depression and (some) attentional disorders are perhaps examples of conditions that were initially characterized as mental or psychological disorders, but which were later given a neurological or neurochemical explanation.

track of information in the system. As noted above, positing suitably interpreted data structures and a read/write memory allows us to keep track of the information conveyed by the bee's dance, that is, the direction, distance, and quality of the nectar. But is that how the bee's nervous system does it? We have granted that implementation-level theories that posit structures that explicitly represent the information attributed to the system at the cognitive level are explanatorily *transparent* relative to that level. But why think that explanatory relationships between levels in science will be transparent? If they were, then science would be much easier than it is. There is no reason to think that Mother Nature, in designing cognitive systems, has been kinder to the theorist of cognition than she has been to the physicist or the biologist.

References

Bard, J. B., & Rhee, S. J. (2004). Ontologies in biology: Design, applications, and future challenges. *Nature Reviews: Genetics, 5*, 213–222.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1*, 371–394.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.

Buchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modeling and fMRI. *Cerebral Cortex, 7*, 768–778.

Burchfield, J. D. (1990). *Lord Kelvin and the age of the earth*. Chicago: University of Chicago Press.

Frisch, K. J. (1967). *The dance-language and orientation of bees*. Cambridge, MA: Harvard University Press.

Friston, K. J. (2003). Dynamic causal models. In R. S. J. Frackowiak, K. J. Friston, C. Firth, R. Dolan, C. J. Price, S. Zeki, J. Ashburner, & W. D. Penny (Eds.), *Human brain function* (2nd ed.). Academic Press.

Friston, K. J., & Buchel, C. (2000). Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proceedings of the National Academy of Science USA, 97*, 7591–7596.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modeling. *Neuroimage, 19*, 1273–1302.

Gallistel, R. C. (2006). The nature of learning and the functional architecture of the brain. In Q. Jing et al. (Eds.), *Psychological science around the world, Vol. 1, Proceedings of the 28th international congress of psychology* (pp. 63–71). Sussex: Psychology Press.

Hamzei, F., Rijntjes, M., Dettmers, C., Glauche, V., Weiller, C., & Buchel, C. (2003). The human action recognition system and its relationship to Broca's area: An fMRI study. *Neuroimage, 19*, 637–644.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences, 1*, 215–226. Reprinted in J. Haugeland (Ed.), (1981) *Mind design* (pp. 243–281). Cambridge, MA: MIT Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Judson, H. (1980). *The eighth day of creation*. New York: Simon and Schuster.

Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2006). Dynamic causal modeling for fMRI: Extending the generative model. Presented at the Rutgers Conference on Philosophical Foundations of Neuroimaging, April 2006. <http://www.philosophy.rutgers.edu/EVENTS/NEUROIMAGING/kiebel1.pdf>

Li, W., Piech, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience, 7*, 651–657.

Lomber, S. G., Payne, B. R., Hildetag, C. C., & Rushmore, J. (2002). Restoration of visual orienting into a cortically blind hemifield by reversible deactivation of posterior parietal cortex or the superior colliculus. *Experimental Brain Research, 142*, 463–474.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77*, 24–42.

Manjaly, Z. M., Marshall. J. C., Stephan, K. E., Gurd, J. M., Zilles, K., & Fink, G. R. (2003). In search of the hidden: An fMRI study with implications for the study of patients with autism and with acquired brain injury. *Neuroimage, 19*, 674–683.

Marr, D. (1974). The computation of lightness by the primate retina. *Vision Research, 14*, 1377–1388.

Marr, D. (1982). *Vision*. New York: Freeman.

Marr, D., & Nishihara, K., (1978). Visual information processing: Artificial intelligence and the sensorium of sight. *Technology Review*, October, 28–48.

Poldrack, R. A., Parker, D. S., & Bilder, R. M. (2006). How can neuroimaging relate cognitive and neural processes? Presented at the Rutgers Conference on Philosophical Foundations of Neuroimaging, April 2006. http://www.philosophy.rutgers.edu/EVENTS/NEUROIMAGING/poldrack.pdf

Sprague, J. M. (1966). Interaction of cortex and superior colliculus in mediation of visually guided behavior in the cat. *Science, 153*, 1544–1547.

van Gelder, T. (1997). Dynamics and cognition. In J. Haugeland (Ed.), *Mind design* (pp. 421–450). Cambridge, MA: MIT Press.